# Automatic Assessment of Children's Reading with the FLaVoR Decoding Using a Phone Confusion Model

*Emre Yılmaz, Joris Pelemans and Hugo Van hamme*

Dept. ESAT, KU Leuven, Belgium

{emre.yilmaz,joris.pelemans,hugo.vanhamme}@esat.kuleuven.be

## Abstract

Reading skills of children can be improved with the help of automatic reading tutors (ART), i.e. interactive software with an appealing interface which supports and challenges the child in the reading task, provides instantaneous feedback and automatically assesses its reading skills. For this purpose, ARTs benefit from automatic speech recognition technology for tracking the child's responses and detecting reading miscues (errors). In previous work, a novel speech recognition architecture has been proposed which adopts a two-layered structure: first a phone recognizer uses task-independent acoustic and language models to generate a phone lattice which is then decoded using a lexicon of expected words and task-dependent finite state grammars. This approach has shown significant improvements in reading miscue detection. In this paper, we extend this technique by employing a more flexible decoding scheme that allows substitution, deletion and insertion of phones. Specifically, the phone lattice generated in the first layer is extended based on a phone confusion matrix that models the typical phone confusions in a language. The proposed system has provided improved miscue detection on the CHOREC database compared to a baseline system without a phone confusion model.

**Index Terms**: automatic reading assessment, flavor decoding, reading miscue detection, children's speech, automatic speech recognition

## 1. Introduction

Automatic speech recognition (ASR) technology has been recently used as a part of automatic reading tutors (ART) for assessing and improving the reading level of elementary school children [1–5]. ARTs are also used in the diagnosis and treatment of reading difficulties such as dyslexia. With the advances in ASR and since the conventional methods require considerable time and effort, ARTs have been becoming more viable in recent years. Moreover, automated reading assessment performed by ARTs does not suffer from observer bias which is a serious problem in conventional methods.

The ASR component of an ART processes the child's response to evaluate how well the child articulates the text appearing on the screen. The processing involves tracking the child's reading position and assessing if each word is pronounced correctly. Although the words appearing on the screen are known in advance, this recognition task is challenging due to several reasons. Firstly, children's speech has increased variation in its spectro-temporal content compared to adults' mainly due to poor articulation capabilities, higher fundamental frequencies and lower speaking rates [6–8]. Furthermore, disfluencies such as hesitation, repetition, stuttering are inherent in read speech especially in the early years of elementary school. As a result, the recognition accuracy of ASR systems using acoustic models trained on fluent adult speech is significantly reduced [9, 10]. This performance gap between speech recognition with child versus adult input can be narrowed by training age-dependent acoustic models and applying speaker adaptation techniques [8, 11, 12].

The SPACE project[1] aims at utilizing ASR technology at schools in Flanders targeting pupils aged 6 to 12 years [4, 13]. For this purpose, a two-pass recognition architecture [14, 15], namely the FLaVoR (Flexible Large Vocabulary Recognition) approach, has been adopted in which a phone lattice is generated during the first pass of the decoder using general acoustic and phone language models only. Then the task-specific information is added during the second pass of the decoder in the form of a finite state transducer containing the correct phonetic transcription of the words along with garbage loops to account for unexpected pronunciations. This approach has provided significant improvements in reading miscue detection compared to a single-layered approach using the same resources for lattice generation and decoding [16].

In this paper, we enrich the lattice generated in the first layer using a phone confusion matrix which models the likely phone substitutions, deletions and insertions in a language. In practice, the lattice is extended according to the information in the phone confusion matrix and the extended lattice is decoded using the task-specific resources. A similar setup has been applied on a Dutch large vocabulary continuous speech recognition task and some improvement in the recognition accuracy has been reported in [17] compared to a FLaVoR setup without using a phone confusion matrix. We also expect this flexible decoding scheme to cope better with the challenging nature of child speech, thus provide improved automatic reading assessment.

The rest of the paper is organized as follows. Section 2 gives a detailed explanation of the ASR component including the FLaVoR approach. Section 3 discusses the experiments performed to investigate the performance of the proposed system. The results are presented and discussed in Section 4. Section 5 concludes the paper.

## 2. Recognition and Assessment of Children's Speech

In this section, we first give an overview of the ASR component as illustrated in Figure 1. Further details can be found in [14]. Then, we focus on the structure of finite state transducers (FST) and describe how reading assessment is performed based on the word-level recognition output.

[1] SPeech Algorithms for Clinical and Educational applications. http://www.esat.kuleuven.be/psi/spraak/projects/SPACE

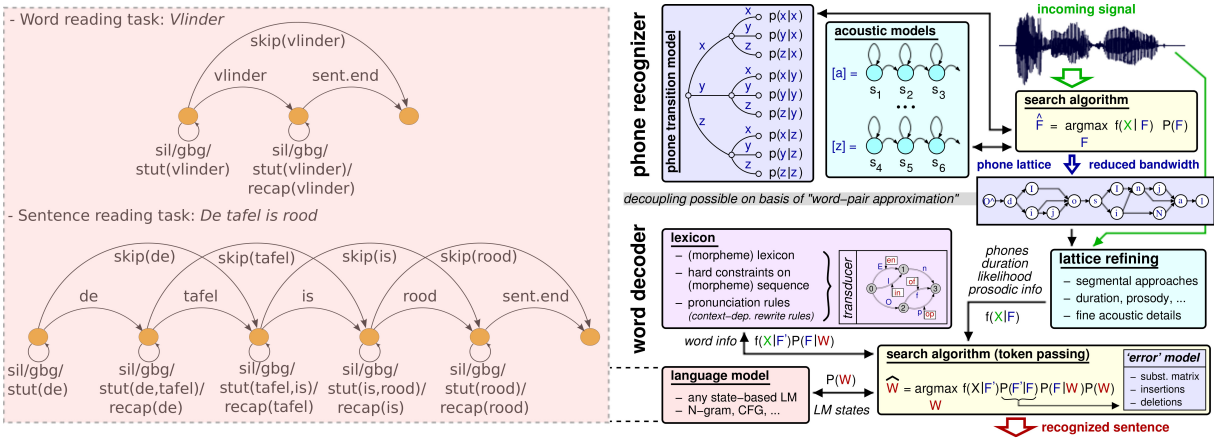


Figure 1: Recognizer overview

### 2.1. The FLaVoR approach

A two-layered HMM-based recognition system is used for obtaining the word-level recognition output. In the first layer, a phone recognizer determines the network of most probable phone strings $F$ (henceforth the phone lattice) given the acoustic features $X$ of the incoming signal. The employed resources are an acoustic model $p(X|F)$ and a phone transition model $p(F)$, which are task-independent models trained on a large database of the target language. The density of the phone lattice is adjusted with predefined beam width thresholds, defined both in terms of a maximum number of active HMM states and a minimum likelihood with respect to the best scoring hypothesis.

The aim of the second layer is to find the most likely phone sequences through the phone lattice generated in the first layer by mapping phone sequences onto the expected words. This is achieved by using task-dependent resources, i.e. a lexicon containing the phonetic transcription of the expected words and a language model in the form of an FST. Since the FST cannot recover from the over-aggressive acoustic pruning in the first layer, a phone confusion model (also called an error model) is applied during lattice decoding. In practice, we train a phone confusion matrix based on a large corpus to find typical confusions, i.e. typical mistakes the recognizer makes, by comparing the output of the recognizer with the transcription [18]. Confusions that are very common, e.g. substituting the two fricatives /v/ and /f/, are expected to have a low cost while very unlikely confusions, e.g. substituting the vowel /a/ and the consonant /d/, are expected to be very costly. For every possible insertion, deletion and substitution (called *phone operations* in the sequel), a cost is obtained to end up with a full confusion matrix containing all the costs.

The phone confusion model is applied to extend the phone sequence hypotheses in the phone lattice during the search process [18]. Several constraints are imposed to limit the number of phone operations. Firstly, each phone operation is penalized with a cost based on the phone confusion matrix. Secondly, a single error constraint is set to prevent the excessive extension of the lattice: after each phone operation, the next phone is required to be correct. By allowing only a single error in a row, the recognized word sequence cannot deviate too much from the phone sequence hypotheses in the lattice. Finally, a pruned version of the confusion matrix, which is obtained by ignoring the phone operations that have a very high cost compared to a threshold value, is used in previous applications [17]. This kind of thresholding also reduces the computational burden at the expense of fewer possible phone operations. Furthermore, from reading assessment point of view, limiting the number of allowed phone operations is crucial as it prevents overdetection of the target word's phonetic transcription in the lattice which masks reading miscue detection.

### 2.2. Reading assessment with FSTs

As the sentences that the child is reading are known, the search space is constrained using a task-specific FST combining a sentence-level FST and a garbage model. Examples of these task-specific FSTs for a word reading task and a sentence reading task are given on the left side of Figure 1. The first example shows how the Dutch word *Vlinder* (Butterfly) is modeled in an FST structure. The second example is a more general one with the Dutch sentence *De tafel is rood* (The table is red).

Three possible reading miscues are modeled in the FSTs shown in Figure 1. Skipping a word (and uttering the next word) is achieved through arcs ending up in the state corresponding to the next word. Skipping multiple words or skipping backwards in time, which is not shown in the figure for the sake of simplicity, can be trivially implemented with the same motivation. Recapping a word is modeled with self-transition arcs. Stuttering arcs cope with partially read words and causes a self-transition like recapping. All of these arcs are penalized with costs to suppress their inappropriate use. Alternative pronunciations of a word are listed in the lexicon and they are accepted as correct pronunciations. In this way, the system can handle pronunciation variations due to speaking style or different dialects.

The main motivation is to explicitly model both the correct pronunciation and all expected, frequent reading miscues, e.g. skipping, recapping and stuttering. Moreover, a general phone model (the garbage model) is also adopted in order to match any unexpected speech. To avoid the frequent use of the garbage model, the garbage loops are penalized with a garbage model cost that is incurred once upon entry.

Using an FST structure that also models the frequent reading miscues has several advantages [15]. Firstly, it provides improved reading assessment with a better detection of the modeled miscues. Moreover, it provides information about the reading miscue type which is needed for appropriate feedback to the child and teacher.

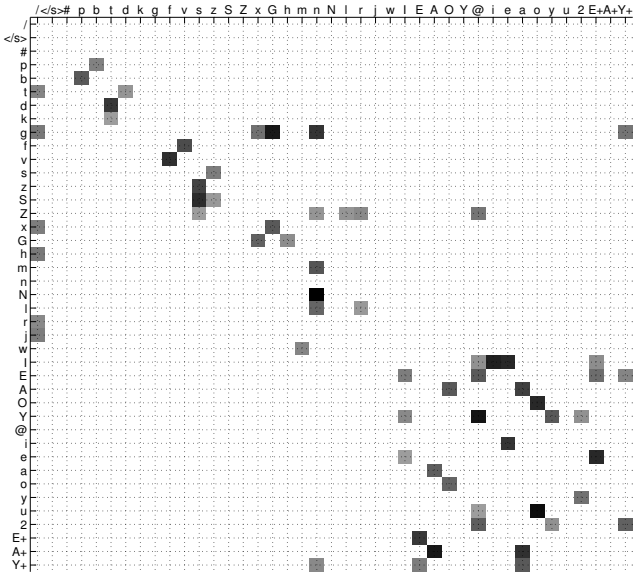


Figure 2: The pruned general phone confusion matrix

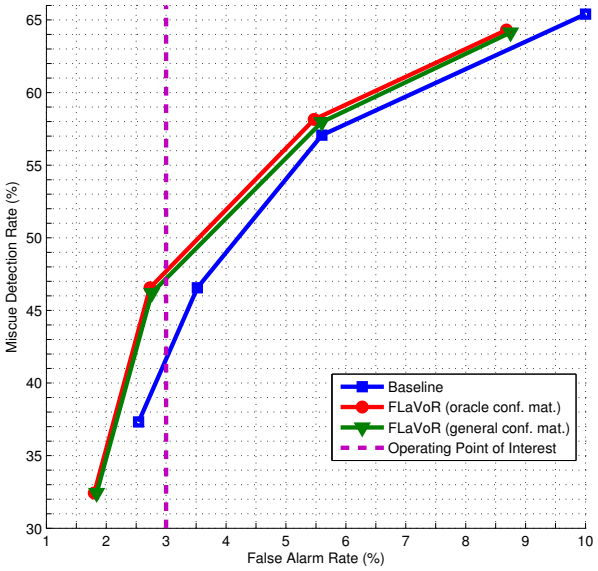


Figure 3: ROC curves obtained using a baseline system without a phone matrix confusion and the proposed system using an oracle and a general confusion matrix

# 3. Experiments

## 3.1. Database and phone confusion models

We have performed the recognition experiments on the CHOREC database [19] which consists of reading sessions from 400 Dutch speaking elementary school children aged between 6 and 12. The reading material contains two word-level reading tests with real and pseudo-words and a story reading task. The word-level reading tasks consist of 3 lists of 40 words with 1, 2 and 3-4 syllables. The recordings were manually segmented, transcribed and annotated with various information such as the target words, phonetic transcription of the utterances and reading miscues.

For training a *general* phone confusion matrix as described in [18], we use a set of recordings consisting of 134 reading sessions (in total 5,360 words), each containing 40 isolated 2-syllable real words. The recognition experiments have been performed on a different set of recordings with 182 reading sessions of the same reading task (in total 7280 words) read by different children. To investigate how well the *general* phone confusion model generalizes to different speakers, we have also trained an *oracle* phone confusion matrix on the same 182 recordings. Thresholding is applied to both confusion matrices to obtain the pruned versions with a limited number of allowed phone operations.

The pruned version of the general phone confusion model is illustrated in Figure 2. The darker the boxes, the higher the corresponding phone operation costs. In this figure, the YAPA phonetic symbol set is used to represent the phonetic symbols [20]. # and </s> refer to silence and sentence end respectively. From this figure, it can be seen that the lattice is extended using this phone confusion matrix during decoding so that it is able to handle several common phone confusions in Dutch, such as /p/-/b/, /t/-/d/, /f/-/v/, /m/-/n/ and /G/-/x/.

## 3.2. Speech recognition system

The first layer uses acoustic models that are trained on a Dutch corpus different from the CHOREC database, containing 22 hours of spoken or read children speech aged between 5 to 11. The cross-word context dependent acoustic models have 1343 tied HMM states and 16,054 tied Gaussian distributions in total (on average 94.6 Gaussians per state).

The preprocessing step involves calculation of the mel spectrum, vocal tract length normalization [15], cepstral mean normalization and discriminative linear feature transformation [21]. The frame length is 25 ms and the frame shift is 10 ms. Further details of the preprocessing can be found in [4]. The general language models used in the first layer are trigram phoneme sequence model estimated from a large Dutch database with correctly read sentences [4].

## 3.3. Evaluation metrics

For each of 40 2-syllable words to be read from the screen, the novel automatic reading assessment system has to perform a binary classification (either correctly pronounced or mispronunciation). We evaluate the accuracy of the classification using the receiver operating characteristic (ROC) curve. The evaluation is performed based on the final trial of each word, i.e. a word is labeled as correctly pronounced if the child pronounces it correctly at the final attempt, otherwise it is a mispronunciation.

We use two well-known measures, namely miscue detection rate (true positive rate) and false alarm rate (false positive rate), to compare the performances of the proposed scheme with the baseline system which does not use a phone confusion model. Miscue detection rate (MDR) is the percentage of the reading miscues that are correctly classified as a mispronunciation by the reading tutor. False alarm rate (FAR) is the percentage of the correctly pronounced words that are incorrectly classified as a mispronunciation by the reading tutor.

# 4. Results and Discussion

In this section, we compare the ROC curves of three automatic reading assessment system, i.e. the baseline system which does

not use a phone confusion model and two novel systems using the oracle and general phone confusion models. The ROC curves obtained for each system is given in Figure 3. The child error rate, i.e. the total percentage of reading errors made by the children, is equal to 7.6%. The different operating points on the figure are obtained by adjusting the density of the lattice generated in the first layer. Denser lattices lead to less detected reading miscues due to the increased probability of finding the correct phonetic transcription in the lattice.

From the ROC curves, it can be concluded that the novel reading assessment system using a phone confusion model outperforms the baseline system, especially at lower FARs. We choose a FAR value of 3% as a feasible operating point to compare the MDR provided by each system. At this FAR value, the impact of recognizer errors on the reading assessment is assumed to be negligible as the child error rate of 7.6% is significantly larger than the chosen FAR value. The baseline system provides a MDR of 41.7% at the FAR value of 3%. The novel systems using the oracle and general phone confusion matrices improve the MDR significantly ($p < 0.01\%$) to 47.7% and 47.2% respectively.

These results illustrate the effectiveness of the flexible decoding scheme with a phone confusion model in the automatic reading assessment task. Moreover, there is only a small gap between the performance of the oracle and general phone confusion matrices. Hence, possible phone confusions in a reading task modeled by the confusion matrix are mostly independent from the speaker identity and a confusion matrix trained on a certain task is expected to generalize well to different recordings with similar speech material.

## 5. Conclusions

Automatic reading tutors are valuable tools for assessing children's reading levels in a repeatable and consistent manner. By assessing children's reading level automatically, a significant amount of human effort can be focused on other educational tasks. However, this is a difficult task in practice due to the high spectral and temporal variation that is inherent to children's speech.

This paper describes a novel automatic reading assessment system which combines a two-layered recognition architecture with a phone confusion model. This phone confusion model allows a limited amount of phone substitution, insertion and deletion resulting in a more flexible phone lattice decoding.

The recognition results show that using the system including a phone confusion model performs better miscue detection with an absolute improvement of 5.5%. Furthermore, a phone confusion model trained on the development data brings similar improvements compared to a confusion model trained on the test data. From this result, it can be concluded that a phone confusion model generalizes well to recordings uttered by different speakers.

## 6. Acknowledgments

This work has been supported by the KU Leuven research grant OT/09/028 (VASI) and IWT-SBO Project 100049 (ALADIN).

## 7. References

[1] J. Mostow, S. Roth, A. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proc. of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, USA, 1994, pp. 785–792.

[2] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, St. Thomas, USA, 2003, pp. 186–191.

[3] X. Li, L. Deng, Y. C. Ju, and A. Acero, "Automatic children's reading tutor on hand-held devices," in *Proc. INTERSPEECH*, Brisbane, Australia, September 2008, pp. 1733–1736.

[4] J. Duchateau, Y. O. Kong, L. Cleuren, L. Latacz, J. Roelens, A. Samir, K. Demuynck, P. Ghesquire, W. Verhelst, and H. Van hamme, "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules," *Speech Communication*, vol. 51, no. 10, pp. 985–994, October 2009.

[5] P. Price, J. Tepperman, M. Iseli, T. Duong, M. P. Black, S. Wang, C. K. Boscardin, M. Heritage, P. D. Pearson, S. S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Communication*, vol. 51, no. 10, pp. 968–984, Oct. 2009.

[6] S. Eguchi and I. J. Hirsch, "Development of speech sounds in children," *Acta Oto-Laryngol. Suppl.*, vol. 257, pp. 1–51, 1969.

[7] S. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of children's speech: Developmental changes in temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, pp. 1455–1468, 1999.

[8] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, Nov. 2003.

[9] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, Washington, DC, USA, 1996, pp. 349–352.

[10] M. Russell and S. D'Arcy, "Challenges for computer recognition of childrens speech," in *Speech and Language Technology in Education (SLATE)*, Farmington, PA, USA, 2007, pp. 108–111.

[11] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. ICASSP*, Seattle, WA, USA, 1998, pp. 433–436.

[12] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, Oct. 2007.

[13] L. Cleuren, "Elements of speech technology based reading assessment and intervention," Ph.D. dissertation, KU Leuven, October 2009.

[14] K. Demuynck, T. Laureys, D. Van Compernolle, and H. Van hamme, "FLavor: a flexible architecture for LVCSR," in *Proc. EUROSPEECH*, Sept. 2003, pp. 1973–1976.

[15] J. Duchateau, M. Wigham, K. Demuynck, and H. Van hamme, "A flexible recognizer architecture in a reading tutor for children," in *Proc. of the ITRW on Speech Recognition and Intrinsic Variation*, Toulouse, France, May 2006, pp. 330–331.

[16] J. Duchateau, K. Demuynck, and H. Van hamme, "Evaluation of phone lattice based speech decoding," in *Proc. INTERSPEECH*, Brighton, UK, Sept. 2009, pp. 1179–1182.

[17] J. Pelemans, K. Demuynck, and P. Wambacq, "A layered approach for Dutch large vocabulary continuous speech recognition," in *Proc. ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4421–4424.

[18] K. Demuynck, D. Van Compernolle, and H. Van hamme, "Robust phone lattice decoding," in *Proc. ICSLP*, 2006, pp. 1622–1625.

[19] L. Cleuren, J. Duchateau, P. Ghesquire, and H. Van hamme, "Children's oral reading corpus: Description and assessment of annotator agreement," in *Proc. 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 2008.

[20] P. Mertens and F. Vercammen, "Fonilex manual," Tech. Rep., March 1998.

[21] K. Demuynck, "Extracting, modelling and combining information in speech recognition," Ph.D. dissertation, K.U.Leuven, ESAT, February 2001.